DE GRUYTER MOUTON
DOI 10.1515/cllt-2013-0009 — Corpus Linguistics and Ling. Theory 2013; 9(1): 1–38

Philip Durrant

Formulaicity in an agglutinating language: the case of Turkish

Abstract: This study examines the extent to which complex inflectional patterns found in Turkish, a language with a rich agglutinating morphology, can be described as formulaic. It is found that many prototypically formulaic phenomena previously attested at the multi-word level in English – frequent co-occurrence of specific elements, fixed 'bundles' of elements, and associations between lexis and grammar – also play an important role at the morphological level in Turkish. It is argued that current psycholinguistic models of agglutinative morphology need to be complexified to incorporate such patterns. Conclusions are also drawn for the practice of Turkish as a Foreign Language teaching and for the methodology of Turkish corpus linguistics.

Keywords: formulaic language, collocation, collostruction, lexical bundles, usage-based model, Turkish, morphology

Philip Durrant: Graduate School of Education, Bilkent University, Ankara, Turkey
E-mail: durrant.phil@googlemail.com

1 Introduction

The study of formulaic language is based around the insight that some linguistic sequences which could potentially be analyzed into smaller units are, for one reason or another, better treated as wholes (Durrant & Mathews-Aydınlı, 2011). In some cases, sequences need to be treated as wholes because their meaning or syntactic behaviour is not predictable from a more general knowledge of the language. Examples include idioms (e.g. *the last straw*), opaque collocations (e.g. *French windows*), and the 'formal idioms' discussed within construction grammar (e.g. *the –er the –er*) (Fillmore, Kay, & O'Connor, 1988). In other cases, sequences are treated as wholes because, although they are semantically and syntactically regular, they have been accepted by the speech community as the usual way of expressing a particular message. Examples include phrases which have become linked to particular contexts (e.g. *long live the king; as shown in Table . . .*) and transparent collocations (e.g. *answer the phone; commit a crime*). Because the adoption of one form rather than another is largely arbitrary, nativelike production

requires specific knowledge of such forms (Pawley & Syder, 1983). Finally, a sequence may be considered a formula if it occurs so frequently that some form of independent storage in long-term memory is cognitively more efficient than creating the sequence from scratch each time it is needed (Goldberg, 2006, p. 64).

The study of formulaicity is closely associated with usage-based models of language (Kemmer & Barlow, 2000). According to such models, a speaker's language system is intimately bound up with their lifetime's experience of the language. This is perhaps most prominently seen in the strong relationships which are held to exist between the frequencies of occurrence of various aspects of the language and their representation and processing by native speakers. Ellis (2002, 2008) has documented these relationships at length, showing how implicit learning mechanisms 'tune' the language system, creating sensitivity to frequency of occurrence across all linguistic levels. The psycholinguistic evidence reviewed by Ellis shows frequency to affect the processing of phonology, phonotactics, reading, spelling, lexis, morphosyntax, formulaic language, language comprehension, grammaticality, sentence production, and syntax.

According to usage-based models, formulaicity can emerge in various ways (Ellis, 2003): through regular association between particular complex forms and particular contexts, leading to those forms' entrenchment as formulaic items; through regular co-occurrence of words (or other linguistic units), leading to their mutual association, and hence status as collocations which have psychological reality for native speakers; and through the grammar learning process, in which even the most general syntactic representations emerge only from a gradual process of abstraction from lexically-specific exemplars, and never entirely lose their association with those concrete forms (Kemmer & Barlow, 2000, p. ix). On this view, the dichotomy between abstract syntax and concrete vocabulary – or between rules and lists – is considered to be a false one. There is, rather, a continuum between wholly memorized and wholly rule-based constructions, with most forms falling somewhere between these extremes (Langacker, 1987). Hence, apparently abstract syntactic constructions may be associated with particular lexis (Hoey, 2005; Hunston & Francis, 2000; Stefanowitsch & Gries, 2003) and apparently memorized forms (such as idioms) may be subject to syntactic processing and variation (Gibbs, Nayak, & Cutting, 1989; Peterson, Burgess, Dell, & Eberhard, 2001).

The formulaic perspective on language has provided important insights in a wide range of areas, including theoretical linguistics, psycholinguistics, corpus linguistics, second language learning, and natural language processing (see Wray, 2008 for a recent overview). A weakness of work in this area at present, however, is its focus on a rather narrow range of, usually European, languages, and especially on English. This has meant both that the benefits of taking a for-

mulaic approach to language have been restricted to these languages and that the status of formulaicity as a general principle of language (rather than a quirk of a few selected languages) remains insufficiently firmly established. It is therefore important that formulaic language research be broadened to a wider range of languages.

As Biber (2009) has recently noted, a particularly interesting area for exploration is that of agglutinating languages, such as Turkish and Finnish, which make use of extensive systems of suffixes to build up complex word forms. The study of formulaicity in such languages is interestingly different from that in English in that their rich morphology raises the possibility that complex types of formulaicity may take place within, as well as between, formulaic words. Definitions of formulaicity have long acknowledged that formulaic language can include linguistic units at all levels (e.g., Wray, 2002, p. 9), and there is a rich psycholinguistic literature on the respective roles of memory and rules in the processing of morphologically complex words (see, for example, the papers in Baayen & Schreuder, 2003). However, there has been little consideration of how morphological formulaicity might function in agglutinating languages. While some work has been done on the holistic processing of morphologically complex words in Finnish (Section 2, below), we shall see that these studies have adopted a rather simple dualistic model on which words are either stored as a fixed wholes or processed morpheme-by-morpheme. I will argue that, given the complexity of morphology in agglutinating languages, such an all-or-nothing dichotomy may be too simple to capture the full range of formulaic morphology in such languages. Formulaic morphology, I will claim, is likely to include patterns falling between the extremes of full-form storage and full morphemic processing. Hypotheses about the types of formulaic patterns which might exist within morphologically complex words cannot be based on psycholinguistic data alone; they will require detailed corpus-based descriptions of the repeated patterns found in such languages.

The primary aim of the present paper is to provide an initial description of such patterns for Turkish. In particular, it will consider the extent to which three widely-researched formulaic phenomena – syntagmatic association between items (as in collocation), fixed sequences of items (as in lexical bundles), and associations between particular lexical and grammatical forms (as in collostruction) – are demonstrated at the morphological level in Turkish. Each of these types of formulaicity offers a distinct, but incomplete, viewpoint on the repetitive patterning which exists in a language. Since these categories of formula were developed in the contexts of other languages, they may, ultimately, not be the best approaches to capturing Turkish formulaicity. However, it will be seen that the combination of the distinct viewpoints offered by each phenomenon both allows an evaluation of competing models of morphology and gives pointers

to ways in which the study of formulaic patterning in Turkish morphology might be further developed.

As well as extending our knowledge of formulaic language in general, and Turkish morphology in particular, studying formulaic morphology in Turkish will also, it is hoped, have more applied benefits. Descriptions of formulaic phenomena in English have provided important bases for applications such as dictionary writing, language pedagogy, and natural language processing, as well as serving as a foundation for the development of many corpus-linguistic methodologies. It is hoped that a description of formulaicity in Turkish may provide similar benefits for that language. This is especially important at the present moment in time, as corpus-based work in the language seems likely to accelerate rapidly in the coming years, with the recent release of the first Turkish National Corpus (www.tnc.org.tr).

2 Formulaicity in agglutinating languages

Turkish, like Hungarian and Finnish, is an agglutinating language, building up sometimes extremely complex word forms through an extensive range of suffixes. Though the distinction can be a problematic one (Beard, 1998), grammars traditionally divide Turkish suffixes into the derivational and the inflectional. Derivation is defined as "the creation of a new lexical item (i.e. a word form which would be found in a dictionary)" (Göksel & Kerslake, 2005, p. 52). Attaching a derivational suffix to a word creates a new word related in meaning to its stem, though the transparency of the connection between stem and word is variable. While a few derivational suffixes are still 'productive', in that they have a regular meaning and can be used with any stems fitting certain criteria, the majority are unproductive – i.e. they can be discerned within already existing words but native speakers no longer perceive them as available for use in the production of new words (Göksel & Kerslake, 2005, p. 52). Inflectional suffixes, on the other hand, are perceived as productive. Their primary functions are to indicate the relations between sentence constituents and to mark functional relations such as case, person, and tense (Göksel & Kerslake, 2005, p. 68).

We can take as a simple example of inflection the word *olabileceğini* (attested 18 times in newspaper corpus described below), which is found in contexts such as:

(1a)	*kasetin*	*doğru*	***olabileceğini***	*düşünüyor*
	cassette-GEN	genuine	**be-POSS-SUB-POS.3-ACC**	think-PROG.3

believe(s) that the cassette may be genuine

This word comprises the root form *ol* ('be') and four suffixes:

(1b)	*ol*	*POSS-<y>Abil*	*SUB-AcAK*	*POS.3-<s>I<n>*	*ACC-<y>I*
	root	suffix 1	suffix 2	suffix 3	suffix 4

The notation used here indicates both the function and the phonemic form of the suffix (see Appendix for a full list of suffixes and their functions). The first suffix indicates possibility, the second indicates subordination through nominalization, the third indicates third person singular possession and the forth shows that the form is in the accusative case. It should be noted that Turkish morphemes are subject to a number of regular phonemic rules which can alter their phonological realisations. In the notation used here, and throughout the present paper, angled brackets indicate that a letter is included only if the suffix is adjacent to a vowel, while capitalisation indicates that a letter changes according to context, for example to ensure vowel harmony. In the present example, the *A*s in suffixes 1 and 2 become *E*s and the *K* in suffix 2 is softened to *Ğ*.

There has to date been little investigation of the role of formulaicity in Turkish. Tannen and Öztürk (1989) and Doğançay (1990) describe the use and extent of situational formulas in conversation, and report such formulas to be extremely common. A few studies have also investigated collocation in Turkish: Oflazer et al. (2004) discuss how an automated corpus parser might handle multiword expressions, Özkan (2007) draws on a corpus of literary works to identify and describe verb-adverb collocations, while Pilten (2008) uses a historical corpus to identify how the collocations of words within a certain semantic set have developed over time. Doğruöz and Backus (2009) consider how Turkish as it is spoken by immigrants in the Netherlands is affected by Dutch formulas. However, none of these studies has addressed the key issue of formulaic patterns in morphology.

The most thorough investigation to date of formulaic phenomena in agglutinating morphology is found in the psycholinguistic literature on the processing of complex words in Finnish. Studies employing a wide range of methodologies (e.g, Lehtonen, et al., 2007; Lehtonen & Laine, 2003; Niemi, Laine, & Tuomainen, 1994; Soveri, Lehtonen, & Laine, 2007; Vartiainen, et al., 2009) have suggested a model on which most inflected nouns are processed morpheme-by-morpheme in both comprehension and production. This is revealed by consistently slower reaction times in lexical decision and naming tasks for inflected than for matched monomorphemic words, by the greater difficulties experienced by an aphasic patient in reading aloud inflected forms, by the same aphasic's errors in producing multimorphemic forms, and by evidence from magnetoencephalography of stronger and longer-lasting activation of the left superior temporal cortex when processing inflected forms. Very high-frequency inflected nouns were immune to

all of these effects, suggesting that such forms are holistically stored. However, it is notable that the frequency level at which holistic storage seems to occur in Finnish is much higher than that previously found for English. Whereas Alegre and Gordon (1999) found evidence for holistic processing in inflected forms with frequencies of around 6/million words, similar effects are seen in Finnish only at frequencies of around 100/million (Soveri, et al., 2007). It may be therefore that holistic storage is rarer in agglutinating languages than in languages with less rich morphologies.

While these studies provide a large amount of important evidence, one potential shortcoming is that they work within a dualistic paradigm, on which complex words are held to be either stored whole or fully processed. This model is at odds with usage-based models' rejection of the dichotomy between lexis and syntax, and two different strands of research suggest that it may be too simplistic to capture the full extent of formulaicity. Firstly, within studies of formulaic language, corpus evidence suggests that formulas rarely consist of fully fixed strings. There is, rather, a predominance of partially-fixed sequences and probabilistic associations between linguistic units (e.g., Biber, 2009; Cheng, Greaves, & Warren, 2006; Moon, 1998; Sinclair, 2004). Psycholinguistic studies of idiom processing have shown that this flexibility within formulaic sequences is reflected in processing, with the moveable semantic sub-parts of idioms playing an important role in their processing (Gibbs, et al., 1989).

The other strand of research which suggests that a whole-word vs. morpheme dichotomy may be too simplistic comes from connectionist models of morphology. According to such models, the psychologically-real components of words emerge from a language user's input on the basis of overlaps between the words they encounter. Thus, a word-part such as past tense *-ed* is psychologically real only to the extent that it receives analogical support from its appearance across a range of words. Research in this tradition has shown that partial matches between words, which include not only word parts which are themselves words (e.g. the *walk* in *walked*) and traditional morphemes (e.g. the *ject* in *inject* and the *ed* in *walked*) but also phonaesthemes (e.g. the *fl* in 'liquid' words, such as *flow, float, flood*) can prime words with overlapping forms and meanings (Hay & Baayen, 2005).

Both of these lines of research suggest that a model on which the only alternative to full morpheme processing is whole word retrieval may be too blunt to reflect the true nature of formulaic morphology. In order to develop more fine-grained models of what might be formulaic in agglutinating morphology, it will be necessary to provide a thorough corpus-based description of the repetitive patterns of morphology in agglutinating languages. The primary aim of the present paper is to provide an initial description of such patterning in Turkish.

3 Methodology

3.1 Corpus

Following researchers in the Firthian tradition (e.g., Hoey, 2005; Sinclair, 2004; Stubbs, 1996), formulaicity is defined here in terms of high frequency of occurrence in a corpus. High-frequency forms are of interest in themselves in terms of what they can tell us about the nature of discourse (as discussed, for example, in the work of Stubbs (1996) and Biber and his colleagues (Biber, Conrad, & Cortes, 2004; Biber, Johansson, Leech, Conrad, & Finegan, 1999)). However, the central focus of the present study – like that of Hoey (2005) – is on the hypotheses that high-frequency patterns can suggest regarding the nature of psycholinguistic processing and representation.

As with Hoey's work, the inference from corpus to mind relies on two key assumptions. The first is the usage-based principle that the frequency with which features of the language are experienced in input influences the representation of those features in the language system. The weight of psycholinguistic evidence in this area (see, e.g., Ellis, 2002, 2008) makes this assumption look plausible. However, it should be borne in mind that links between frequency and representation are complex and affected by a variety of other factors (Ellis & Larsen-Freeman, 2006). Frequency-based hypotheses about the likely psycholinguistic realisation of any patterns found must therefore be read as exploratory and provisional. Provided this caveat is kept in mind, though, corpus analysis of this sort has the potential to open up new avenues of psycholinguistic research whose potential may not become clear from online studies alone.

The second key assumption is that the corpus investigated is representative of the input experienced by language users. Unfortunately, this assumption is often an unjustified one. Corpora are usually intended as balanced samples of texts from across a particular language domain, whether that domain be a national variety (e.g. British National Corpus, Corpus of Contemporary American English, METU Turkish corpus) or a more specialised field of discourse (e.g. Michigan Corpus of Academic Spoken English, British Academic Written Corpus). While such corpora are excellent for some purposes, none of them resemble any language user's likely experience with the language: no one will be exposed to the full range of different styles present in the British National Corpus or to the range of academic topics and genres covered by the Michigan Corpus (Hoey, 2005, p. 14).

This mismatch between the range of texts usually found in corpora and what a language user actually experiences is a serious shortcoming of previous work in

this area. This is acknowledged by Hoey (2005), who emphasises that his research into lexical priming is not able to indicate the primings of any particular language user. However, he maintains that corpora can indicate "the kinds of data a language user might encounter" and so suggest "the ways in which priming might occur and the kinds of feature for which words or word sequences might be primed" (Hoey, 2005, p. 14). No argument is presented for this claim, however, and it is not clear what Hoey's optimism is based on. There is every reason to believe that the highly skewed language exposure which is likely to be typical of most language users is radically different from that which is found in balanced corpora. Discourse analysts would be rightly distrustful of any research which tried to draw conclusions about 'academic language' based on a corpus of one individual's exposure to the domain. We should be equally suspicious of research drawing conclusions about individuals' exposure based on broadly representative corpora. This point is especially important in research on formulaic language, where the natural skew inherent in an individual's experience with the language is likely to be an important factor in increasing the formulaicity of their input (c.f., Goldberg, 2006).

With this in mind, the current study does not draw on existing large-scale corpora of Turkish. It uses instead a corpus representing one language user's (the present author's) exposure to a single discourse type (online newspapers) in Turkish over a period of 6 months. Like many foreign/second language learners, my main source of extensive reading in my target language is online newspapers. Between November 20, 2009 and May 20, 2010, I stored everything I read in this form as text files, and collated these files to create the corpus. Over the six-month period, I read on a total of 111 days (usually reading on five days each week) and for a mean of 50 minutes per day. In total, I read 765 texts, or part texts[1] – 515 news items and 250 opinion pieces – totalling 374,590 words. Though these texts were taken from 7 different newspapers, the vast majority were taken from a single title. Table 1 shows the total number of texts of each type taken from each newspaper.

Table 1: Contents of the newspaper corpus (numbers of articles per publication)

Title	Radikal	Taraf	Vatan	Milliyet	Zaman	Cumhuriyet	Hürriyet	Total
News items	450	39	13	4	4	4	1	515
Opinion pieces	174	48	6	13	9	0	0	250

1 Only those parts of the text which I actually read were recorded.

Clearly, given the nature of the corpus, the degree to which specific patterns found can be generalised to other language users and other text types is an empirical question, open to further investigation. What we can say, however, is that, given such experience, these patterns are likely to be real for this particular reader with regard to this particular discourse type. Moreover, if we make the highly plausible assumption that this reader's experience is at least analogous to that of other language users, the *types* of frequency-based biases seen here are likely to be typical of any individual user's input.

3.2 Procedure

Whereas work on suffixation in Finnish has focused on suffixes following nouns, the present research will look instead at those following verbs since the latter comprises a much more extensive morphological set and so a richer database for study. Twenty verbs will be analysed. It seems likely *a priori* that formulaicity will be more prominent in higher than in lower frequency verbs. Verbs were therefore selected to cover a wide range of different frequencies. The verbs, their English translations, and their frequencies in the corpus are shown Table 2.

All inflected forms of these verbs were retrieved from the corpus using an alphabetical word list generated by WordSmith Tools (Scott, 2008) and each form was manually coded for inflectional suffixes. This coding was based on the list of inflectional suffixes provided by Göksel and Kerslake (2005). All suffixes identified with these verbs are shown with their frequencies in the corpus and brief characterisations of their functions in the Appendix.

Two important points should be noted here. First, the analysis considers only inflectional morphemes. This focus was chosen because inflectional forms are considered to be, on the whole, actively productive. The question of formulaicity is, therefore, more open than for derivations, which, as we have seen, are considered to be largely unproductive (though it is worth noting that the Finnish morphology studies suggest that derived forms are holistically stored in the input lexicon only, with morpheme-based processing employed in production (Laine, Niemi, Koivuselkä-Sallinen, Ahlsén, & Hynönä, 1994)). Second, some distinct morphemes in Turkish are orthographically indistinguishable from each other. For example, the subordinator *SUB-<y>AcAK* is identical to the future tense marker *FUT-<y>AcAK*. In such cases, concordance lines were consulted to determine the frequency of each morpheme.

All inflected forms of the twenty verbs under investigation were listed in a spreadsheet, along with their frequencies of occurrence, with each column in the spreadsheet representing one suffix. Table 3 illustrates the contents of the

Table 2: Verb stems studied

Verb root	Translation	Cumulative stem frequency[2]	Total types	% total tokens covered by top 5% of types	% types appearing once only
ol	be	8,540	438	72.06	39.27
et	do/make	4,161	423	56.50	42.08
yap	do/make	3,189	355	57.54	42.54
ver	give	1,836	256	49.35	39.45
de	say	1,232	108	60.71	44.44
çık	go/come out; emerge	1,112	145	53.15	42.76
çalış	work	964	167	38.90	43.11
konuş	speak	790	157	53.29	50.96
geç	pass	768	188	43.88	54.26
yaşa	live/experience	736	156	45.92	51.92
gir	enter/go into	474	133	38.19	55.64
bak	look	381	111	37.27	55.86
bırak	leave	341	114	31.67	56.14
anlat	explain	313	80	48.56	50.00
geliş	develop	312	70	37.82	51.43
sağla	provide/obtain	271	97	32.47	56.70
yarat	create	207	73	37.20	60.27
koru	protect	190	64	43.68	64.06
paylaş	share	76	41	18.42	68.29
önle	prevent	42	22	21.43	68.18

Table 3: Sample analysis

Word	Freq.	Root	Suffix 1	Suffix 2	Suffix 3
Anlatmaması	1	anlat	*NEG-mA*	*SUB-mA*	*POS.3-<s>I<n>*
Anlatmıyor	1	anlat	*NEG-mA*	*IMP-<I>yor*	
Anlatmadım	1	anlat	*NEG-mA*	*PRF-DI*	*1-m*
Anlatın	3	anlat	*2PL-<y>In*		
Anlatsın	1	anlat	*3-sIn*		

2 In this paper, the term 'cumulative stem frequency' is used to refer to the combined frequency of all inflected forms of a verb.

spreadsheet. This spreadsheet formed the database for all of the analyses which follow. All subsequent analyses were performed on the basis of this spreadsheet, using the open-source statistical package *R* (R Core Development Team, 2010).

4 Results

4.1 Overview of form frequencies

As Table 2 shows, a large number of differently inflected forms (*types*) were found for each of the 20 verbs studied. The frequencies of these forms, unsurprisingly, follow Zipf-like skewed distributions, with a very small number of high-frequency forms accounting for a high percentage of each verb's tokens and a long 'tail' of very low-frequency forms accounting for a high percentage of their types.

These distributions suggest that both formulaicity and productive inflection are common. The ten most frequent forms identified all occurred on average more than once in every twenty minutes of reading time, strongly suggesting some type of formulaicity. At the other end of the spectrum, 39–68% of types of each lemma studied were hapax logomena, suggesting high levels of productivity (Baayen, 2008). The primary aim of this study, however, was to consider whether more complex formulaic patterns can be found below the level of the word. We shall now turn to this issue

4.2 Collocation between suffix combinations

The aim of this section is to determine the extent to which high-frequency combinations of morphemes occur across words. The initial approach will be to start with individual suffixes and determine the extent to which they enter into regular relationships with one another. On analogy with corpus work on relations between orthographic words, I will refer to such relationships as *collocations* between suffixes.

We can get an initial impression of the nature of collocation between suffixes by looking in detail at a few examples. Tables 4–7 show all of the morphological environments in which the four suffixes with the highest type frequencies (i.e. those used in the greatest number of different words) in the corpus appear and the percentage of tokens of the node suffix which are found with each collocate (I will take this percentage to indicate the *strength* of a collocational relationship). The columns in these tables indicate the position of the collocates – L1 collocates

Table 4: Collocations of the 'NEG-mA' morpheme (The percentage of occurrences of the node with which each morpheme occurs in this position is given in brackets)

L3	L2	L1	Node	R1	R2	R3
CAUS-DIr (0.14%) CAUS-Ir (0.09%) SUB-<y>Iş (0.05%)	PASS-Il (3.73%) CAUS-DIr (0.32%) PASS-<I>n (0.28%) SUB-<y>Iş (0.23%) CAUS-Ir (0.14%)	POSS-<y>A (18.60%) PASS-Il (6.68%) PASS-<I>n (0.92%) CAUS-DIr (0.46%) CAUS-Ir (0.28%) SUB-<y>Iş (0.28%) CAUS-t (0.28%)	NEG-mA	SUB-DIK (19.80%) AOR-z (17.31%) SUB-<y>An (10.13%) PRF-DI (10.08%) IMP-<I>yor (9.53%) SUB-<y>AcAK (6.22%) SUB-mA (5.52%) FUT-<y>AcAK (5.16%) EV/PRF-mIş (4.14%) COND-sA (1.75%) SUB-mAK (1.52%) 3-sIn (1.52%) OBLG-mAlI (1.24%) 2PL-<y>In (1.15%) 1PL-<y>Iz (0.97%) 1-m (0.69%) POSS-<y>Abil (0.64%) SUB-<y>IncA (0.46%) IPV-mAktA (0.37%) SUB-<y>ArAk (0.37%) SUB-DIKchA (0.32%) OPT-<y>A (0.32%) 3PL-sIn<lAr> (0.18%) SUB-<y>Ip (0.18%)	POS.3-<s>I<n> (26.38%) 3PL-<lar> (3.36%) PRF-DI (2.44%) POS.3PL-lArI<n> (2.12%) 1PL-<y>Iz (2.07%) COP-DIr (1.98%) COND-sA (1.93%) 1-<y>Im (1.52%) PL-lAr (1.15%) 1PL-k (1.01%) Pos-1pl<I>mIz (0.78%) 2PL-sInIz (0.74%) 1-m (0.69%) SUB-<y>ken (0.55%) COP-y (0.51%) AOR-<A/I>r (0.46%) CIC-<y>lA (0.41%) ACC-<y>I (0.37%) POS.1-<I>m (0.32%) 2-sIn (0.32%) 1PL-lIm (0.23%) EV/PRF-mIş (0.18%) GEN-<n>In (0.14%) SUB-<y>AcAK (0.14%) DAT-<y>A (0.14%) POS.2PL-<I>nIz (0.14%) 2F-nIz (0.14%) POS.2-<I>n (0.09%) 1-yIm (0.09%) ABL-DAn (0.05%) 2-n (0.05%) SUB-DIK (0.05%)	ACC-<y>I (12.48%) DAT-<y>A (1.66%) GEN-<n>In (1.1%) PRF-DI (0.69%) ABL-DAn (0.55%) COND-sA (0.23%) POS.3-<s>I<n> (0.18%) Gen-DIr (0.18%) 1-m (0.14%) LOC-DA (0.14%) 1PL-k (0.14%) CIC-<y>lA (0.14%) 2F-nIz (0.09%) 1PL-<y>Iz (0.09%) COP-y (0.05%) EV/PRF-mIş (0.05%)

Table 5: Collocations of the 'SUB-mA' morpheme (The percentage of occurrences of the node with which each morpheme occurs in this position is given in brackets)

L3	L2	L1	Node	R1	R2	R3
SUB-<y>lş (0.42%)	PASS-Il (1.1%)	PASS-Il (13.92%)	SUB-mA	POS.3-<s>l<n> (49.37%)	ACC-<y>I (9.12%)	AP-ki<n> (0.26%)
PASS-Il (0.06%)	CAUS-DIr (0.81%)	NEG-mA (3.87%)		DAT-<y>A (10.92%)	DAT-<y>A (6.83%)	PRF-DI (0.16%)
CAUS-Ir (0.03%)	SUB-<y>lş (0.61%)	PASS-<I>n (2.45%)		POS.3PL-lAr<n> (6.64%)	GEN-<n>In (6.06%)	ACC-<y>I (0.06%)
	CAUS-Ir (0.48%)	CAUS-DIr (1.58%)		PL-lAr (6.25%)	LOC-DA (2.67%)	COND-sA (0.03%)
	POSS-<y>A (0.26%)	POSS-<y>AbIl (1.19%)		ACC-<y>I (4.35%)	ABL-DAn (2.51%)	
	PASS-<I>n (0.06%)	SUB-<y>lş (0.81%)		GEN-<n>In (3.25%)	CIC-<y>IA (1.71%)	
	COP-DIr (0.03%)	CAUS-Ir (0.55%)		POS.1PL-<I>mIz (1.97%)	COP-DIr (0.77%)	
	CAUS-t (0.03%)	CAUS-t (0.16%)		POS.1-<I>m (1%)	POS.1PL-<I>mIz (0.19%)	
				LOC-DA (0.97%)	COP-y (0.19%)	
				POS.2PL-<I>nIz (0.26%)	PRF-DI (0.03%)	
				CIC-<y>IA (0.19%)	AP-ki<n> (0.03%)	
				COP-DIr (0.13%)		
				POS.2-<I>n (0.1%)		
				COP-y (0.03%)		

Table 6: Collocations of the 'PASS-Il' morpheme (The percentage of occurrences of the node with which each morpheme occurs in this position is given in brackets)

L3	L2	L1	Node	R1	R2	R3
	SUB-<y>Iş (1.93%)	CAUS-DIr (3.74%) CAUS-Ir (1.93%) CAUS-t (0.1%) PASS-<I>n (0.07%) COP-DIr (0.03%)	PASS-Il	SUB-<y>An (23.65%) SUB-mA (14.41%) PRF-DI (11.17%) SUB-DIK (8.81%) EV/PRF-mIş (6.24%) IMP-<I>yor (6.1%) SUB-<y>AcAK (5.6%) NEG-mA (4.84%) FUT-<y>AcAK (4.1%) POSS-<y>Abil (3.34%) AOR-<A/I>r (3.34%) POSS-<y>A (2.7%) SUB-<y>ArAk (1.27%) 3-sIn (0.8%) SUB-mAK (0.77%) OBLG-mAlI (0.57%) IPV-mAktA (0.53%) SUB-<y>Ip (0.5%) SUB-mAdAn (0.33%) COND-sA (0.3%) SUB-<y>IncA (0.2%) SUB-<y>Iş (0.17%) SUB-mAksIzIn (0.13%) SUB-<y>IncAyA (0.13%)	POS.3-<s>I<n> (23.48%) PRF-DI (2.74%) NEG-mA (2.7%) COP-DIr (2.57%) AOR-<A/I>r (1.43%) COND-sA (1.37%) SUB-<y>AcAK (1.33%) SUB-mA (1.13%) SUB-<y>ken (1%) DAT-<y>A (0.97%) AOR-z (0.83%) PL-lAr (0.77%) SUB-DIK (0.77%) EV/PRF-mIş (0.47%) IMP-<I>yor (0.43%) FUT-<y>AcAK (0.4%) SUB-<y>An (0.37%) 3PL-<lar> (0.37%) POS.3PL-lArI<n> (0.33%) ABL-DAn (0.3%) COP-y (0.23%) OBLG-mAlI (0.2%) ACC-<y>I (0.1%) GEN-<n>In (0.1%) CIC-<y>lA (0.07%) 2PL-sInIz (0.07%) POS.1PL-<I>mIz (0.07%) IPV-mAktA (0.03%) POS.2-<I>n (0.03%) 1-m (0.03%) SUB-<y>IncA (0.03%)	ACC-<y>I (6.2%) POS.3-<s>I<n> (2.47%) DAT-<y>A (1.73%) GEN-<n>In (1.6%) AOR-z (1.17%) LOC-DA (0.97%) ABL-DAn (0.83%) PRF-DI (0.5%) SUB-<y>AcAK (0.47%) COND-sA (0.33%) SUB-<y>An (0.27%) COP-DIr (0.27%) CIC-<y>lA (0.2%) SUB-DIK (0.2%) IMP-<I>yor (0.17%) SUB-<y>ken (0.13%) EV/PRF-mIş (0.13%) POS.3PL-lArI<n> (0.1%) 3PL-<lar> (0.07%) SUB-mA (0.07%) 2PL-sInIz (0.03%) COP-y (0.03%) 1-m (0.03%) SUB-DIKchA (0.03%) 3-sIn (0.03%) FUT-<y>AcAK (0.03%)

Table 7: Collocations of the 'POS.3-<s>I<n>' morpheme (The percentage of occurrences of the node with which each morpheme occurs in this position is given in brackets)

L3	L2	L1	Node	R1	R2	R3
PASS-Il (1.39%)	PASS-Il (13.21%)	SUB-DIK (61.23%)	POS.3-<s>I<n>	ACC-<y>I (30.46%)	AP-ki<n> (0.11%)	
POSS-<y>A (1.2%)	NEG-mA (10.73%)	SUB-mA (28.74%)		DAT-<y>A (6.08%)	PRF-DI (0.11%)	
CAUS-DIr (0.71%)	POSS-<y>Abil (2.44%)	SUB-<y>AcAK (9.77%)		GEN-<n>In (3.51%)	COND-sA (0.02%)	
SUB-<y>Iş (0.53%)	PASS-<I>n (2.25%)	SUB-<y>An (0.15%)		LOC-DA (2.53%)		
CAUS-Ir (0.43%)	CAUS-DIr (1.09%)	SUB-<y>Iş (0.09%)		ABL-DAn (1.95%)		
PASS-<I>n (0.21%)	SUB-<y>Iş (0.58%)	AOR-z (0.02%)		CIC-<y>IA (0.88%)		
NEG-mA (0.08%)	CAUS-Ir (0.28%)			COP-DIr (0.41%)		
COP-DIr (0.02%)	CAUS-t (0.04%)			COP-y (0.13%)		
CAUS-t (0.02%)	AOR-<A/I>r (0.02%)					

are those one position to the left of the node, R2 collocates are those two positions to the right, and so on.

These data suggest a number of working hypotheses. First, it seems that the strongest collocates are usually found immediately adjacent to the node morpheme (i.e. at L1 and R1). The only exception is seen in Table 4, where the strongest right-hand collocate of *NEG-mA* is *POS.3-<s>I<n>*, found at R2. This is likely to be a product of the large number of subordinating morphemes (i.e. those prefixed *SUB-*) found at R1 (a similar effect also seen in Table 6, where again the *POS.3-<s>I<n>* suffix is common at R2). The possessive *POS.3-<s>I<n>* is (as Table 7 demonstrates) commonly used directly after such morphemes – a relationship analogous to possessive + gerund forms in English. Thus, while most important collocational relations seem to hold between directly adjacent morphemes, there also seem to be some discontinuous collocations, recalling the lexical bundles with variable slots identified in English by Biber (2009).

Second, all four morphemes appear in pairings which occur once only, even at the immediately adjacent L1/R1 slots, where collocational relations tend to be strongest: *CAUS-t* + *NEG-mA* (Table 4); *SUB-mA* + *COP-y* (Table 5); *COP-DIr* + *PASS-Il* (Table 6); and *AOR-z* + *POS.3-<s>I<n>* (Table 7) are all hapax logomena. All four node morphemes therefore take part in novel (or at least very low-frequency) combinations. At the same time, all four morphemes also enter into strong collocations. Almost a third of cases of *NEG-mA* are directly preceded by the morpheme *POSS-<y>A* (Table 4), while well over a third of cases of *SUB-mA* are directly followed by *POS.3-<s>I<n>* (Table 5). Most strikingly of all, over 99% of occurrences of *POS.3-<s>I<n>* are directly preceded by one of just three other morphemes (Table 7). *PASS-Il* forms weaker relationships, but nevertheless in around one in four occurrences it is directly followed by *SUB-<y>An* (Table 6).

The description so far makes clear that there is considerable variation between suffixes in the types and strengths of collocational relationship into which they enter. Before drawing any strong conclusions, therefore, it will be important to extend the analysis beyond this small set. To get a broader picture, I calculated for all high-frequency morphemes (i.e., the 29 morphemes which appear in 100 distinct words or more), the total number of morphemes appearing in each position from L3 to R3. To estimate the predictability of their collocational environments, I also calculated the percentage of cases of the node morpheme occurring with one of the top 3 collocates in that position (or fewer if the node did not have as many as 3 collocates in that position. For example, only one suffix is ever found directly after *POSS-<y>A*, so the 100% figure reported for that morpheme is based on this one collocate only). This percentage will be referred to as a morpheme's *collocate predictability*. Table 8 shows the results for each morpheme; Table 9 summarises the collocate predictability at each position.

Table 8: Number and strength of collocating morphemes for high-frequency suffixes (In each collocate cell, the first number indicates the total number of different morpheme types found in this position; the second number indicates the percentage of occurrences of the node with which the three (or fewer) most frequent morpheme types appear)

Row no.	L3	L2	L1	Node (type/token frequency)	R1	R2	R3
8.1	3 / 0.28%	5 / 4.33%	7 / 26.2%	NEG-mA (711/2,172)	24 / 47.24%	32 / 32.18%	16 / 15.24%
8.2	9 / 3.3%	9 / 26.37%	6 / 99.74%	POS.3-<s>I<n> (608/5,331)	8 / 40.05%	3 / 0.24%	0 / 0%
8.3	0 / 0%	1 / 1.93%	5 / 5.77%	PASS-Il (544/2,998)	24 / 49.23%	31 / 28.92%	26 / 10.41%
8.4	3 / 0.52%	8 / 2.51%	8 / 20.24%	SUB-mA (528/3,103)	14 / 66.94%	11 / 22.01%	4 / 0.48%
8.5	9 / 1%	20 / 4.69%	19 / 20.03%	PRF-DI (424/4,118)	8 / 12.94%	1 / 0.41%	3 / 0.15%
8.6	4 / 0.27%	8 / 1.83%	10 / 18.49%	SUB-DIK (398/4,040)	9 / 94.65%	11 / 39.48%	4 / 0.15%
8.7	10 / 23.43%	12 / 90.3%	9 / 92.74%	ACC-<y>I (296/2,134)	0 / 0%	0 / 0%	0 / 0%
8.8	4 / 1.88%	7 / 10.6%	8 / 45.89%	SUB-<y>AcAK (284/1,009)	7 / 64.12%	6 / 35.98%	1 / 0.1%
8.9	2 / 0.52%	5 / 5.34%	9 / 25.73%	IMP-<I>yor (260/1,772)	9 / 24.45%	9 / 1.05%	0 / 0%
8.10	0 / 0%	4 / 0.95%	7 / 21.99%	POSS-<y>Abil (248/632)	15 / 78.01%	18 / 28.32%	8 / 9.97%
8.11	2 / 0.15%	6 / 4.89%	7 / 31.43%	AOR-<A/I>r (240/1,330)	11 / 38.57%	9 / 4.51%	1 / 0.08%
8.12	3 / 0.59%	7 / 1.92%	10 / 28.37%	SUB-<y>An (220/3,899)	10 / 11.62%	11 / 5.36%	5 / 0.21%
8.13	6 / 1.02%	11 / 2.77%	13 / 27.17%	EV/PRF-mIş (216/1,082)	8 / 45.01%	5 / 4.53%	0 / 0%
8.14	1 / 0.25%	3 / 1.47%	5 / 22.11%	POSS-<y>A (202/407)	2 / 100%	19 / 63.88%	20 / 26.78%
8.15	4 / 0.71%	5 / 3.46%	8 / 26.63%	FUT-<y>AcAK (195/984)	9 / 30.79%	8 / 1.63%	0 / 0%
8.16	0 / 0%	0 / 0%	2 / 51.86%	CAUS-DIr (191/403)	22 / 53.1%	37 / 28.04%	17 / 14.14%
8.17	3 / 2.58%	5 / 9.78%	5 / 91.44%	POS.3PL-lArI<n> (177/736)	8 / 43.48%	2 / 0.41%	0 / 0%
8.18	8 / 3.06%	12 / 17.01%	16 / 43.71%	COND-sA (176/588)	6 / 21.26%	1 / 10.2%	2 / 1.7%
8.19	9 / 13.01%	13 / 50.88%	8 / 90.28%	DAT-<y>A (172/792)	0 / 0%	0 / 0%	0 / 0%
8.20	6 / 8.41%	9 / 22.63%	11 / 71.55%	3PL-<lar> (162/464)	5 / 13.15%	1 / 0.22%	0 / 0%
8.21	7 / 3.48%	14 / 29.51%	17 / 72.34%	COP-DIr (154/488)	3 / 0.82%	1 / 0.2%	1 / 0.2%
8.22	0 / 0%	0 / 0%	0 / 0%	PASS-<I>n (136/559)	22 / 55.64%	23 / 32.56%	15 / 10.38%
8.23	2 / 0.51%	6 / 16.64%	5 / 99.49%	PL-lAr (132/583)	12 / 45.8%	7 / 2.06%	0 / 0%
8.24	8 / 15.52%	9 / 71.26%	8 / 81.61%	GEN-<n>In (128/522)	1 / 0.38%	0 / 0%	0 / 0%
8.25	4 / 4.07%	7 / 21.29%	8 / 87.56%	1PL-<y>Iz (112/418)	1 / 0.24%	0 / 0%	0 / 0%
8.26	1 / 0.08%	5 / 1.83%	7 / 8.04%	SUB-mAK (106/1256)	4 / 16.16%	1 / 0.24%	0 / 0%
8.27	7 / 13.06%	10 / 45.54%	10 / 83.12%	ABL-DAn (106/314)	2 / 1.27%	0 / 0%	0 / 0%
8.28	0 / 0%	0 / 0%	1 / 1.35%	SUB-<y>Iş (105/370)	19 / 79.73%	26 / 31.08%	34 / 15.41%
8.29	3 / 2.32%	9 / 12.25%	4 / 92.72%	POS.1PL-<I>miz (101/302)	6 / 21.19%	0 / 0%	0 / 0%

Table 9: Summary of collocation strengths

	L3	L2	L1	R1	R2	R3
Median	1.45	7.56	37.57	40.05	5.36	1.7
Min	0.08	0.95	1.35	0.24	0.2	0.08
Max	23.43	90.3	99.74	100	63.88	26.78
Interquartile range	3.11	19.72	62.15	39.715	29.27	12.1

As Tables 8 and 9 show, collocate predictability does, in general, drop rapidly as collocates move further from the node. Only four morphemes have a figure of greater than 45% at L2, i.e.: *ACC-<y>I* (Row 7); *DAT-<y>A* (Row 19); *GEN-<n>In* (Row 24); *ABL-DAn* (Row 27).

The occurrence of these discontinuous collocates appears to be closely related to forms' grammatical function. These four morphemes (Accusative, Dative, Genitive, and Ablative suffixes) correspond to four of the five suffixes classified by Göksel and Kerslake as 'Case Suffixes' (Göksel & Kerslake, 2005, p. 70), used here to modify nominalized verbs. The fifth case suffix, *LOC-DA* (not listed in Table 8 because it did not meet the frequency requirements) also shows high restriction at L2, with 85.65% of cases featuring one of three morphemes.

Only one suffix has similarly high levels of association at R2. This is the suffix *POSS-<y>A*, which indicates impossibility and is always followed by the negative suffix *NEG-mA*. The former is exceptional in being the only morpheme examined to be always accompanied by one particular collocate (suggesting, as we shall see below, that these two items may constitute a single complex unit). Given the strength of its R1 collocation, it is unsurprising that it also has strong collocates at R2.

In the node-adjacent slots (L1 and R1), collocate predictability is generally high – the top 3 (or fewer) collocates accounting for, on average, around 38–40% of occurrences of the node. As the interquartile ranges indicate, however, there is wide variation around this median. Though the majority of suffixes form strong collocations either to left or right – 19/29 suffixes have a score of at least 50% on one side or the other – a few suffixes do not seem to enter into strong collocational relations. In particular, the infinitive form *SUB-mAK* (Row 26) and the perfective *PRF-DI* (Row 5) are both used together with a wide range of other suffixes and do not have any collocate predictability scores exceeding 20%.

At the other end of the scale, the high level of collocate predictability of many morphemes at L1 or R1 suggests that there is likely to be some psycholinguistic link between them and their collocates, or that the combinations may be independently stored items. For 9/29 node suffixes, for example, over 90% of occur-

Table 10: 10 most common adjacent two-word collocations

Bundle	Types	Roots	% verb form tokens
SUB-DIK POS.3-<s>I<n>	197	20	12.59
POS.3-<s>I<n> ACC-<y>I	152	19	6.26
SUB-mA POS.3-<s>I<n>	249	20	5.91
PASS-Il SUB-<y>An	37	15	2.73
SUB-<y>AcAK POS.3-<s>I<n>	152	19	2.01
PASS-Il SUB-mA	99	16	1.81
NEG-mA SUB-DIK	86	18	1.68
POSS-<y>A NEG-mA	200	17	1.56
SUB-DIK POS.3PL-lArI<n>	69	18	1.48
NEG-mA AOR-z	97	17	1.45

rences are found together with one of the top 3 (or fewer) collocates. This hypothesis is reinforced by the very high frequency of some of these pairings. Table 10 shows the 10 most frequent pairings found. Absolute frequency data are not given since the present data can tell us only how often the combinations occurred in inflections of the 20 verbs studied, not how frequent they were in the corpus as a whole. Frequency is therefore given in terms of the percentage of verb tokens which include each pair. To indicate how widely each bundle is used, figures are also given for the number of different verb types (out of the 3,198 forms studied) and the number of different verb roots (out of the 20 studied) with which they were found.

Assuming that the verb roots studied are representative of verbs in general, these pairings are extremely common. The most frequent pairing (the third-person subordinating form *SUB-DIK POS.3-<s>I<n>*) was found in 12.59% of verb forms, suggesting that it must occur at least once in every few lines of text. Indeed, even the least frequent pairing in this top 10 is found on average around once every 70 verbs. These extremely high frequencies, together with the wide spread of the collocations across different verb roots and the regular syntactic function which each performs, suggest that some form of psycholinguistic entrenchment is extremely likely.

From the analysis so far, we can draw two main conclusions. First, most high-frequency suffixes enter into both novel and regular combinations. Thus, while the language system must allow for the use of each individual morpheme in novel constructions, the dominance of certain high-frequency pairings suggests that an efficient usage-based system should also represent the typical immediate morphological environments of many suffixes. Second, strong collocations usually consist of two adjacent morphemes only. The exceptions are combinations which

include a case suffix and those which include the negative possibility marker *POSS-<y>A*. The longer combinations which include these elements will be examined in more detail in the next section.

4.3 Morphemic bundles

To examine in more detail the prevalence and nature of longer collocations, I generated lists of the most common three- and four-morpheme bundles. The term *bundle* is adopted from the work of Biber and his colleagues (1999) on recurrent word combinations, and is used here to refer to frequently recurring sequences of morphemes.

The ten most common combinations of each length are shown in Tables 11 and 12. Examples from the corpus of each suffix cluster in use are also given, along with an English gloss. It should be noted these clusters can be integrated with other morphemes to make still longer words, but that for simplicity of presentation, such examples are not included in these tables. As in Table 10, frequency is given in terms of the percentage of verb tokens which include each bundle and the number of different verb roots with which they were found. For ease of presentation, bundles will be referred to using the row numbers given in the tables.

These data reveal a number of important facts. First, the three-morpheme combinations in Table 11 are all very common indeed. Since they also have regular grammatical functions, this makes some kind of holistic storage seem likely, given a usage-based view of language. Assuming that the verbs studied here are representative of those in the rest of the corpus, the most common combination (11.1) is used in almost one in twenty verb tokens. A speaker who had memorised all of the top ten three-morpheme forms would on average meet one of them in 12.38% of the verbs found in the present corpus. The four-morpheme combinations are considerably less frequent, between them accounting for only 2.22% of verbs.

Second, most of the three-morpheme combinations are used across a wide range of verb roots, all but two (11.8, 11.10) being attested with at least three quarters of the verbs studied. This reinforces the impression that an efficient language system may draw on some kind of representation of these bundles for use across a range of lexical contexts. For four-morpheme suffixes, the spread is much narrower, the majority of combinations being used with fewer than half of the twenty verbs studied.

Third, frequent combinations of three or more items are dominated by two structural types. 18/20 of the combinations identified include combinations of

Table 11: 10 most common three-morpheme bundles

Row no.	Bundle	Examples	Types	Roots	% verb form tokens
11.1	SUB-DIK	"Çocuk-lar-a okul yetiştir-e-mi-yor-uz" de-**diğ-in-i** öne sür-dü	59	17	4.57
	POS.3-<s>I<n>	"child-PL-DAT school keep.up-POSS-NEG-PROG-1PL" say-**SUB-POS.3-ACC** claim-PRF.3			
	ACC-<y>I	claimed that he said that "We can't build enough schools for the children"			
11.2	PASS-Il	Çünkü bırak-**ıl-ma-sı** iste-n-iyor	67	16	1.50
	SUB-mA	Because release-**PASS-SUB-POS.3** request-PASS-PROG.3			
	POS.3-<s>I<n>	Because it is requested that he be released			
11.3	NEG-mA	olumlu bir sinyal ver-**me-diğ-i** muhakkak	51	16	1.48
	SUB-DIK	positive a message give-**NEG-SUB-POS.3** is.certain			
	POS.3-<s>I<n>	it certainly didn't give a positive message			
11.4	PASS-Il	darbe hazırlık-ların-ın konuş-**ul-duğ-u** toplantı-da	46	16	0.97
	SUB-DIK	coup preparation-POS.3PL-GEN discuss-**PASS-SUB-POS.3** meeting-LOC			
	POS.3-<s>I<n>	at the meeting in which the coup preparations were discussed			
11.5	SUB-<y>AcAK	görüşme sürec-in-de bir tıkanıklık ol-**acağ-*ın-ı*** san-mı-yor-um	53	16	0.86
	POS.3-<s>I<n>	meeting process-POS.3-LOC a hold.up be-**SUB-POS.3-ACC** think-NEG-PROG-1			
	ACC-<y>I	I don't think there will be a hold up in the meeting process			
11.6	SUB-mA	Hem de parlemento çalış-**ma-sın-ı** bil-iyor-um	38	18	0.82
	POS.3-<s>I<n>	Also parliament work-**SUB-POS.3-ACC** know-PROG-1			
	ACC-<y>I	And I know the workings of parliament			
11.7	SUB-mA	olay-lar-ın çık-**ma-sın-a** engel ol-ma-ya çalış-tı	27	16	0.61
	POS.3-<s>I<n>	incident-PL-GEN occur-**SUB-POS.3-DAT** obstruction be-SUB-DAT try-PRF.3			
	DAT-<y>A	tried to stop incidents occurring			
11.8	POSS-<y>A	Öğrenci sınıf-a şapka ile gir-**e-me-z**	39	13	0.58
	NEG-mA	Student class-DAT hat-CVB enter-**POSS-NEG-AOR.3**			
	AOR-z	A student cannot go to class wearing a hat			
11.9	SUB-DIK	bir mektup yolla-yarak yaşa-**dık-ların-ı** biz-ler-le paylaş-mak iste-miş-ler	22	15	0.51
	POS.3PL-lArI<n>	a letter send-CVB experience-**SUB-POS.3PL-ACC** us-PL-OBL share-INF want-PRF-3PL			
	ACC-<y>I	by sending a letter they wanted to share what they had experienced with us			
11.10	SUB-DIK	Başkan-ın anlat-**tığ-ın-a** göre	18	12	0.48
	POS.3-<s>I<n>	Chair.perso-GEN explain-**SUB-POS.3-DAT** according.to			
	DAT-<y>A	According to the chair person			

Table 12: 10 most common four-morpheme bundles

Row no.	Bundle	Examples	Types	Roots	% verb form tokens
12.1	NEG-mA SUB-DIK POS.3-<s>I<n> ACC-<y>I	mümkün ol-**ma-dığ-ın-ı** söyle-yerek possible be-**NEG-SUB-POS.3-ACC** say-CVB saying that is was not possible	20	12	0.71
12.2	PASS-Il SUB-DIK POS.3-<s>I<n> ACC-<y>I	böylece kent-ler-de bir tür coşku yarat-**ıl-dığ-ın-ı** in.this.way city-PL-DAT a sort.of excitement create-**PASS-SUB-POS.3-ACC** belirt-erek explain-CVB explaining that a sort of excitement was created in the cities in this way	13	10	0.35
12.3	PASS-Il SUB-mA POS.3-<s>I<n> ACC-<y>I	seçim yap-**ıl-ma-sın-ı** iste-dik-lerin-i söyle-di election do-**PASS-SUB-POS.3-ACC** want-SUB-POS.3PL-ACC say-PRF.3 said that they wanted elections to be held	12	11	0.24
12.4	NEG-mA SUB-<y>AcAK POS.3-<s>I<n> ACC-<y>I	Meclis-'te kal-ma-mız-ın bir anlam-ın-ın Parliament-LOC stay-SUB-POS.3PL-GEN a meaning-POS.3-GEN ol-**ma-yacağ-ın-ı** ifade et-miş-ti-k be-**NEG-SUB-POS.3-ACC** express make-PSTPRF-1PL We had said that there would be no point in our staying in parliament	19	9	0.16
12.5	POSS-<y>A NEG-mA SUB-<y>AcAK POS.3-<s>I<n>	artık böyle ol-**a-ma-yacağ-ı** için bu iş-ler now like.this be-**POSS-NEG-SUB-POS.3** because this event-PL ol-uyor be-PROG.3 these events are occurring because it won't be possible to carry on like this any more	21	11	0.15

12.6	PASS-Il SUB-mA POS.3-<s>I<n> GEN-<n>In	farklılık-lar-ın ifade ed-**il-me-sin-in** düşman-ı difference-PL-GEN expression make-**PASS-SUB-POS.3-GEN** enemy-POS.3 an enemy to differences being expressed	9	9	0.13
12.7	PASS-Il SUB-mA POS.3-<s>I<n> DAT-<y>A	çıplak ol-arak deniz-e gir-**il-me-sin-e** izin naked be-CVB sea-DAT enter-**PASS-SUB-POS.3-DAT** permission ver-il-me-yecek give-PASS-NEG-FUT.3 permission to go into the sea naked will not be given	6	5	0.13
12.8	PASS-Il POSS-<y>A NEG-mA AOR-z	Bu iddia-lar cevap-sız bırak-**ıl-a-ma-z** This claim-PL answer-NEG leave-**PASS-POSS-NEG-AOR.3** These claims cannot be left unanswered	5	4	0.13
12.9	POSS-<y>Abil SUB-<y>AcAK POS.3-<s>I<n> ACC-<y>I	katarakt-tan koru-**yabil-eceğ-in**-i göster-diğ-in-i cataract-ABL protect-**POSS-SUB-POS.3-ACC** show-SUB-POS.3-ACC belirt-ti explain-PRF.3 explained that this showed that it could protect against cataracts	12	8	0.13
12.10	PASS-Il NEG-mA SUB-mA POS.3-<s>I<n>	Böylesi önemli bir tesis-e zarar ver-**il-me-me-si** Such important a facility-DAT harm give-**PASS-NEG-SUB-POS.3** gerek-iyor necessitate-PROG.3 such an important facility shouldn't be harmed	10	6	0.09

Table 13: Relations between bundles

			SUB-DIK	POS.3PL-lArI<n>	ACC-<y>I
			SUB-DIK	POS.3-<s>I<n>	ACC-<y>I
		NEG-mA	SUB-DIK	POS.3-<s>I<n>	
		NEG-mA	SUB-DIK	POS.3-<s>I<n>	ACC-<y>I
		PASS-Il	SUB-DIK	POS.3-<s>I<n>	
		PASS-Il	SUB-DIK	POS.3-<s>I<n>	ACC-<y>I
			SUB-DIK	POS.3-<s>I<n>	DAT-<y>A
			SUB-mA	POS.3-<s>I<n>	ACC-<y>I
		PASS-Il	SUB-mA	POS.3-<s>I<n>	
		PASS-Il	SUB-mA	POS.3-<s>I<n>	ACC-<y>I
			SUB-mA	POS.3-<s>I<n>	DAT-<y>A
		PASS-Il	SUB-mA	POS.3-<s>I<n>	DAT-<y>A
		PASS-Il	SUB-mA	POS.3-<s>I<n>	GEN-<n>In
	PASS-Il	NEG-mA	SUB-mA	POS.3-<s>I<n>	
			SUB-<y>AcAK	POS.3-<s>I<n>	ACC-<y>I
		NEG-mA	SUB-<y>AcAK	POS.3-<s>I<n>	ACC-<y>I
		POSS-<y>Abil	SUB-<y>AcAK	POS.3-<s>I<n>	ACC-<y>I
	POSS-<y>A	NEG-mA	SUB-<y>AcAK	POS.3-<s>I<n>	
PASS-Il	POSS-<y>A	NEG-mA	AOR-z		
	POSS-<y>A	NEG-mA	AOR-z		

subordinators plus person markers, while the remaining two (11.8, 12.8) involve negative forms. Interestingly, Biber and his colleagues have noted that lexical bundles in English often involve parts of embedded complement clauses, such as *I don't know why* or *I thought that* (Biber, et al., 1999, p. 991). The high-frequency morpheme bundles seen here seem therefore to play a similar structural role to many word bundles found in English – i.e. that of anchoring complex sentences. This fits well with the view that one aim of formulaic language is to enable speakers to fluently negotiate utterances which involve more than a single clause (Pawley & Syder, 2000).

Finally, it is important to note that (as their structural similarities would already suggest) there are strong family resemblances across the most frequent morpheme bundles. Table 13 organizes the combinations to highlight these similarities. Looking across the examples here, the impression is not of twenty separate forms, but rather of a smaller number of form sets, or of prototypes permitting predictable variations. The situation is comparable to that noted in English by, for example, Durrant (2009), who points out that variable forms such as *there was/were (no/a) statistically significant difference(s) between/in* appear to be more common than entirely fixed formulas. Given the prominence of such

cases, the idea of independent, fixed complex forms may be a misleading one. It may be better to see formulas rather as variable prototypes, or families of related forms.

4.4 Spread of suffix combinations across verb roots

The fact that the three-morpheme bundles shown in Table 11 are attested across a wide range of lemmas implies that they are relatively productive – i.e. available for use in many lexical contexts. This might suggest that the language system represents such bundles as purely grammatical items, independent of the lexical roots to which they are applied. Such a conclusion would be surprising from a usage-based perspective however. As was noted in the introduction, usage-based models reject any sharp distinction between grammar and lexis, and one of the key findings of recent corpus linguistics has been that particular grammatical forms are commonly associated with particular lexis, and vice-versa (Hoey, 2005; Hunston & Francis, 2000; Stefanowitsch & Gries, 2003).

This phenomenon has been most systematically studied by Stefanowitsch & Gries (2003), who developed the technique of 'collostruction analysis' to quantify such associations. They use Fisher's exact test to determine the probability that there is a relationship of either attraction or repulsion between a particular lexical item and a particular grammatical form. Significant results are taken to indicate either an attraction or repulsion, and the lexis associated with a particular pattern ranked according to their *p*-values.

Using this method, Stefanowitsch and Gries (2003) demonstrate that syntactic structures at a variety of levels of abstraction are biased towards particular lexical instantiations. Unsurprisingly, the slots in relatively specific constructions like (*X* think nothing of V_{gerund}] are relatively limited in the lexis which they accept (the *V* slot being associated with verbs indicating undesirable/risky activities). Perhaps less expectedly though, even relatively abstract grammatical forms, such as the past tense, are significantly attracted to some verbs (such as *be* and *say*) and repelled by others (such as *know* and *do*). Stefanowitsch and Gries do not attempt to explain associations of the latter sort, but point out that their very existence represents a significant problem for models of language in which syntax is strictly separated from lexis.

The final analysis in this paper will adopt Stefanowitsch and Gries's (2003) technique to determine whether the 3-morpheme bundles identified in the previous section are equally likely to appear with all of the verbs they are attested or whether there are significant relationships of attraction/repulsion between

particular bundles and particular verb roots. Like Stefanowitsch and Gries, I will take a prevalence of such associations to suggest that a model on which suffix bundles are represented entirely independently of lexical roots is problematic.

As a first step, we can consider the lexical associations of the most frequent bundle in the corpus – the *SUB-DIK POS.3-<s>I<n> ACC-<y>I* bundle, which was attested with 17 of the 20 verb roots studied. Table 14 shows all roots which are shown by Fisher's exact test to be either attracted to or repelled by the bundle with a significance of $p < .001$. As is immediately obvious from the table, occurrences of this bundle are overwhelmingly dominated by forms using a single root: *ol* ('be'). If uses of the bundle were distributed between the attested roots in proportion to their appearance in the corpus as a whole (the null hypothesis we would expect were grammar and lexis independent of each other), we would expect to find 395 co-occurrences of this root and suffix bundle. However, the actual count is over twice this figure, at 808. Indeed, over 68% of cases of this bundle are instantiated with this particular root (compared to the 33% we would expect given the overall frequency of *ol*). Moreover, the attraction is mutual: the *ol* root shows a strong preference for this particular bundle, with almost 10% of its occurrences being found with this bundle (compared to the 4.6% we would expect given the overall frequency of the bundle).

Unsurprisingly, given this strong skew towards *ol*, many other roots are found to occur far less frequently with the bundle than the null hypothesis would predict. The strongest repulsion is found for the root *yap* ('do'/'make'), which is found in this form on 70 times, compared to the 147 we would expect on the null hypothesis.

This analysis suggests that, while this bundle is available for use across most verb roots, grammar and lexis are not independent of each other. It remains to be seen, however, to what extent such associations exist for other morphemic bundles. Table 15 shows the roots with are attracted to/repelled by each of the ten most frequent three-morpheme bundles with a significance of $p < .001$. The most obvious, and most important, point to note here is that only one bundle (*SUB-DIK POS.3PL-lArI<n> ACC-<y>I*) does not show any such associations. The phenomenon of collostruction therefore appears to be widespread in Turkish. This suggests that lexis and syntax are unlikely to be entirely distinct in the language system of the reader whose experience is represented by the current corpus.

Second, certain commonalities are evidenced across bundles. Specifically:

– All bundles which include the two-morpheme combination *SUB-DIK POS.3-<s>I<n>* and do not include the passive morpheme *PASS-Il* (rows 15.1, 15.3, and 15.10) are attracted to the root *ol* ('be') and repelled by the roots *yap* ('do'/'make') and *et* ('do'/'make'). The *ol* root it not attracted to, and the *et*/*yap* roots not repelled by, any other bundles.

Table 14: Lexical associations of the *SUB-DIK POS.3-<s>I<n> ACC-<y>I* bundle

Relation	Root	Translation	Fisher's test	Frequency		% of bundle		% of root	
				Expected	Actual	Expected	Actual	Expected	Actual
Attracted	ol	be	3.88E-142	395	808	33.3	68.1	4.6	9.5
Repelled	yap	do/make	1.04E-13	147	70	12.4	5.9	4.6	2.2
	et	do/make	1.30E-10	193	115	16.2	9.7	4.6	2.8
	ver	give	2.55E-9	85	37	7.2	3.1	4.6	2.0
	konuş	speak	8.27E-8	37	9	3.1	0.8	4.6	1.1
	de	say	1.73E-6	57	25	7.2	2.1	4.6	2.0
	çık	go/come out; emerge	9.32E-6	51	23	4.3	1.9	4.6	2.1
	anlat	explain	1.43E-5	14	1	1.2	0.08	4.6	0.3
	geç	pass	1.44E-4	36	15	3.0	1.3	4.6	2.0
	çalış	work	2.07E-4	45	22	3.8	1.9	4.6	2.3
	bırak	leave	9.25E-4	16	4	1.3	0.3	4.6	1.2

Table 15: Associations and repulsions between three-morpheme bundles and verb roots

Row no.	Bundle	Attracted			Repelled		
		Root	Translation	Fisher's test (*p*)	Root	Translation	Fisher's test (*p*)
15.1	SUB-DIK POS.3-<s>I<n> ACC-<y>I	ol	be	3.88E-142	yap et ver konuş de çık anlat geç çalış bırak	do/make do/make give speak say go/come out; emerge explain pass work leave	1.04E-13 1.30E-10 2.55E-9 8.27E-8 1.73E-6 9.32E-6 1.43E-5 1.44E-4 2.07E-4 9.25E-4
15.2	PASS-Il SUB-mA POS.3-<s>I<n>	et yap bırak yarat ver	do/make do/make leave create give	2.06E-24 4.57E-15 2.59E-9 6.34E-7 4.37E-4	çık çalış	go/come out; emerge work	1.57E-6 1.8E-4
15.3	NEG-mA SUB-DIK POS.3-<s>I<n>	ol	be	2.11E-62	yap et konuş çalış	do/make do/make speak work	4.67E-8 4.46E-5 1.40E-4 5.67E-4
15.4	PASS-Il SUB-DIK POS.3-<s>I<n>	yap et	do/make do/make	9.89E-23 3.27E-8	çık	go/come out; emerge	4.16E-4

15.5	SUB-<y>AcAK POS.3-<s>I<n> ACC-<y>I	sağla	provide/ obtain	5.77E-4	çalış	work	3.77E-3
15.6	SUB-mA POS.3-<s>I<n> ACC-<y>I	konuş bırak	speak leave	9.08E-6 7.43E-3	ol	be	5.27E-3
15.7	SUB-mA POS.3-<s>I<n> DAT-<y>A	konuş	speak	3.32E-4			
15.8	POSS-<y>A NEG-mA AOR-z	et	do/make	1.67E-5			
15.9	SUB-DIK POS.3PL-lArI<n> ACC-<y>I						
15.10	SUB-DIK POS.3-<s>I<n> DAT-<y>A	ol	be	2.06E-23	et yap	do/make do/make	1.13E-5 8.56E-3

– Both bundles which include the passive marker *PASS-Il* (rows 15.2 and 15.4) are attracted to the roots *yap* ('do'/'make') and *et* ('do'/'make') and repelled by the root *çık* ('go'/'come out'; 'emerge'[3]).
– Both bundles which include the two-morpheme combination *SUB-mA POS.3-<s>I<n>* and do not include the passive morpheme *PASS-Il* (rows 15.6 and 15.7) are attracted to the root *konuş* ('speak'). Indeed, were the analysis extended to the 11th most frequent bundle (*SUB-mA POS.3-<s>I<n> GEN-<n>In*), which also includes this pair, it would been seen that the same root is again attracted. This root is not attracted to any other bundles.

These considerations suggest that some of the associations seen here hold not between the roots and these specific three-morpheme bundles (which are, like the N-grams of traditional corpus linguistics, mere analytical conveniences), but rather between the roots and more abstract grammatical categories. Any full usage-based description of the likely grammar of this language user would therefore need to determine the level of abstraction at which particular lexical-syntactic associations are most salient.

5 General discussion

This study set out to explore the extent and types of high-frequency morphological patterns that can be found in the input of a language user (specifically, the input of a reader of online newspapers over a six month period) in Turkish. As was noted in the introduction, none of the analyses used here can present a complete picture of formulaicity. However, taken together, they do suggest a number of important conclusions.

First, the corpus evidenced both a small number of very high frequency complex words, and a large range of very low frequency words. Such a distribution fits comfortably within the dual-route processing model put forward for Finnish by Niemi et al. (1994). However, formulaic patterning does not stop at the word level. A morphological model which takes seriously the usage-based claim that frequency affects representation will need to take into account the following facts:

3 Note that in Turkish, intransitive verbs can be used in the passive, where it gives the meaning of something being 'generally done'.

- Most high-frequency morphemes enter into strong collocational relations with their syntagmatic neighbours. This is seen especially strongly between directly adjacent morphemes, but certain morpheme types (especially nominal inflections) also form collocations with more distant morphemes.
- A number of high-frequency three-morpheme combinations exist and are used across a wide range of verb roots. These usually serve the function of verb subordination or negation.
- There are strong family resemblances between these morphological combinations, suggesting that they may not be entirely separate entities, but rather collections of related forms, or variations on a prototype.
- High-frequency morpheme bundles are not neutral with regard to lexis. Rather, they form strong relationships of attraction with particular verb roots.

Models such at that of Niemi et al. (1994), in which words are either fully stored or fully processed, seem poorly suited to accounting for these data since they waste the potential processing economies which are found in frequently repeated patterns below the word level. Similarly, any model in which grammar and lexis are treated as strictly independents systems (e.g, Pinker, 1999) is also unlikely to be adequate, given the strong associations between particular verb roots and morpheme collocations.

More consistent with the present data are models in which morphology operates through paradigmatic analogies, based on relations of formal and semantic similarity found in networks of holistically-stored high-frequency words (see Hay & Baayen, 2005 for a review). Sets of related high-frequency morpheme bundles (such as those shown in Table 13) might be seen as inhering in sets of frequent word-form representations which overlap in form and function, with repeated patterns extended by analogy to new forms. Particular repeated patterns (such as the three-morpheme bundles listed in Table 11) might in some cases emerge as independently-represented entities, but would most likely exist within networks of associations with other morphological patterns (as shown in Table 13) and with particular lexical roots (as shown in Tables 14 and 15).

A second, more applied, implication of the present research is that the teaching of Turkish as a Foreign Language – where, the present author can attest from personal experience, fluently combining long suffix chains in speech is a major challenge for new learners – may profit from a mode of presentation in which learners are made aware of certain suffix pairs and triples and their associations with their most frequent lexical roots. Special attention should be given in this context to forms featuring the high-frequency three-morpheme combinations which mark subordination. A good knowledge of the most common instantiations

of such forms should provide an effective basis for later extending use to more complex and unusual cases.

Third, Turkish corpus linguistics also needs to take account of the types of patterning described here. Since the wide range of complex inflections in Turkish means that many word forms appear with very low frequency, thus making generalization at the lexical level difficult, a natural first step in dealing with corpus data in Turkish is to abstract away from morphology by lemmatizing the corpus. Given the range of possible forms, such lemmatization is itself no trivial task, though promising computational methods are now being developed (e.g, Sak, Güngör, & Saraçlar, 2008). However, the patterning demonstrated in the present paper, and in particular the strong associations between particular verb roots and particular combinations of inflections, suggest that lemmatization should be treated with caution, since its indiscriminate use may conceal phenomena of both applied and theoretical importance.

It has been emphasized that the present study is an exploratory one. Accordingly, it raises rather more questions than it resolves. First, while the frequency-based descriptions provided here can enumerate phenomena which a psycholinguistic model of agglutinative morphology might take into account, and so provide grounds for hypothesis formation, they ultimately examine only the product, not the process of linguistic activity. Further, process-oriented, research is needed to examine the psycholinguistic correlates of these findings before any strong claims can be made about the nature of morphological representation and processing in Turkish.

Second, this study has been restricted to the language experience of one particular user within one particular discourse type (written news) over a relatively limited stretch of time. The question of whether, and in what ways, high-frequency morphological patterns are specific to particular discourse types and particular moments in time remains an open one and requires further investigation.

Finally, this study has limited its attention to formulaic patterning within orthographic words. The extent to which formulaic patterning stretches across word boundaries has not been examined. It seems likely that the nature of collocation between words in Turkish will differ from that in English. This can be seen by taking the simple example of lexical bundles. Whereas Biber's recent study of academic English (2009) found 140 4-grams which occurred with a frequency of at least ten per million words, a similar search of a comparable Turkish academic corpus[4] yielded only 18 4-grams meeting the same frequency criteria, 8 of which

4 A 2.5 million word corpus of academic research articles compiled by the author and balanced across the topic areas of Arts and Humanities, Life Sciences, Science and Engineering, Social Sciences – Administrative, and Social Sciences – Psychological.

Table 16: 4-grams in academic Turkish

Original	English translation	Frequency per million words
istatistiksel olarak anlamlı bir	a statistically significant	31
arasında istatistiksel olarak anlamlı	statistically significant between	30
ve buna bağlı olarak	and connected to this	26
arasında anlamlı bir fark	a significant difference between	24
olarak anlamlı bir fark	a significant difference	20
arasında anlamlı bir ilişki	a significant relationship between	19
deney ve kontrol gruplarının	experimental and control groups	17
diğer bir ifade ile	in other words	16
istatistiksel olarak anlamlı fark	statistically significant difference	16
amerika birleşik devletleri nde	in the united states of america	12
geçerlilik ve güvernilirlik çalışması	validity and reliability study	12
başka bir ifade ile	in other words	12
istatistiksel olarak anlamlı düzeyde	at a statistically significant level	12
dsm-iv tanı ölçütlerine göre	according to dsm-iv measures	10
gelişmiş ve gelişmekte olan	developed and developing	10
iki veya daha fazla	two or more	10
arasında anlamlı bir farklılık	a significant difference between	10
kronik obstrüktif akçiğer hastalığı	chronic obstructive lung disease	10

Table 17: Frequent forms of *prove/ kanıtlamak*

English forms	Frequency/m	Turkish forms	Frequency/m
proved	45	kanıtlar	18
prove	43	kanıtlanmış	8
proven	15	kanıtlanmıştır	7
proves	8	kanıtlanmaktadır	6
proving	6	kanıtlamak	4

are variations on a single pattern describing statistically significant relationships (see Table 16).

The reason for this lack of high-frequency bundles is not hard to find. As we have seen, meanings which require multiple word expressions in English can be expressed using a single word in Turkish. This means that individual word forms (lexemes) have, on average, much lower frequencies than similar word forms in English. Table 17 illustrates this point with a key academic word: the verb *to prove* (*kanıtlamak* in Turkish). The academic section of the Corpus of Contemporary American attests 5 forms of this verb, with the frequencies shown in Table 17. The corpus of academic Turkish, in contrast, attests 55 forms of the verb *kanıtlamak*,

of which over two thirds appear less than once per million words. Even the most frequent forms of the verb (shown in Table 17) appear much less often than the English forms.

Since individual word forms are rare, so too are high-frequency word combinations. This raises the question of the level at which collocational relationships are best described in Turkish. If collocations between individual word forms are relatively rare, then it may be that collocation is better described as relationships between lemmas, or between specifiable subsets of a lemma, or even between suffix combinations, abstracted from lexical roots. Determining such relations will be an important pre-requisite for understanding the implications of lemmatisation for Turkish corpus linguistics and will require further empirical investigation.

Appendix

Suffixes in the corpus (codes indicate both function and phonological form)

Suffix	Function	Token frequency	Type frequency
POS.3-<s>I<n>	Possessive 3rd Person singular	5331	608
PRF-DI	Perfective	4118	424
SUB-DIK	Subordinator (verbal noun/participle/converb)	4040	398
SUB-<y>An	Subordinator (participle)	3899	220
SUB-mA	Subordinator (verbal noun/converb)	3103	528
PASS-Il	Passive	2998	544
NEG-mA	Negative	2172	711
ACC-<y>I	Accusative	2134	296
SUB-<y>ArAk	Subordinator (converb)	1778	37
IMP-<I>yor	Imperfective	1722	260
AOR-<A/I>r	Aorist	1330	240
SUB-mAK	Subordinator (verbal noun/converb)	1256	106
EV/PRF-mIş	Evidential/perfective	1082	216
SUB-<y>AcAK	Subordinator (verbal noun/participle/converb)	1009	284
FUT-<y>AcAK	Future tense	984	195
DAT-<y>A	Dative	792	172
POS.3PL-lArI<n>	Possessive 3rd Person plural	736	177
POSS-<y>Abil	Possibility	632	248
COND-sA	Conditional	588	176
PL-lAr	Plural	583	132
PASS-<I>n	Passive	559	136
GEN-<n>In	Genitive	522	128
COP-DIr	Generalizing modality marker	488	154

3PL-<lar>	Person: 3rd Person plural	464	162
1PL-<y>Iz	Person: 1st Person plural	418	112
POSS-<y>A	Possibility (negative forms only)	407	202
CAUS-DIr	Causative	403	191
AOR-z	Aorist (after negative only)	376	97
SUB-<y>Iş	Subordinator (verbal noun)	370	105
SUB-<y>Ip	Subordinator (converb)	356	31
ABL-DAn	Ablative	314	106
POS.1PL-<I>miz	Possessive 1st Person plural	302	101
SUB-<y>ken	Subordinator (converb)	270	46
1PL-k	Person: 1st Person plural	244	70
3-sIn	Person: 3rd Person singular	232	40
LOC-DA	Locative	223	81
1-<y>Im	Person: 1st Person singular	221	59
1-m	Person: 1st Person singular	202	70
CAUS-Ir	Causative	188	78
2PL-<y>In	Person: 2nd Person plural/formal	175	21
OBLG-mAlI	Obligative	161	58
CIC-<y>lA	Comitative/instrumentl/conjunctive	145	51
IPV-mAktA	Imperfective	139	50
OPT-<y>A	Optative	138	29
POS.1-<I>m	Possessive 1st Person singular	118	56
2PL-sInIz	Person: 2nd Person plural/formal	114	43
1PL-lIm	Person: 1st Person plural	114	21
SUB-mAdAn	Subordinator (converb)	110	22
COP-y	Copula	94	50
SUB-<y>IncA	Subordinator (converb)	93	24
2F-nIz	Person: 2nd Person singular familiar	83	33
POS.2PL-<I>nIz	Possessive 2nd Person plural/formal	60	38
CAUS-t	Causative	33	23
2-sIn	Person: 2nd Person singular familiar	24	19
1-yIm	Person: 1st Person singular	24	8
3PL-sIn<lAr>	Person: 3rd Person plural	21	13
SUB-DIKçA	Subordinator (converb)	16	11
POS.2-<I>n	Possessive 2nd Person singular familiar	14	11
AP-ki<n>	Attributive/pronominal	13	12
SUB-mAksIzIn	Subordinator (converb)	13	6
2-n	Person: 2nd Person singular familiar	11	10
COMPD-<y>Iver	Compound (indicating speed)	8	7
SUB-<y>IncAyA	Subordinator (converb)	7	4
2PL-sAnIzA	Person: 2nd Person plural/formal	5	4
2PL-<y>InIz	Person: 2nd Person plural/formal	4	3
SUB-cAsInA	Subordinator (converb)	4	3
SUB-<y>A . . . <y>A	Subordinator (converb)	4	2
COMPD-<y>Agel	Compound (indicating continuous action)	2	2
SUB-<y>AlI	Subordinator (converb)	2	1
COMPD-<y>Adur	Compound (indicating continuous action)	1	1

Acknowledgments

The author would like to thank Garrett Hubing for his suggestions on annotation and two anonymous reviewers for their constructive comments on a previous draft of the paper. Any remaining errors are, of course, the author's responsibility.

References

Alegre, Maria, & Peter Gordon. 1999. Frequency effects and the representational status of regular inflections. *Journal of memory and language* 40(1). 41–61.

Baayen, R. Harald. 2008. Corpus linguistics in morphology: Morphological productivity. In A. Ludeling & M. Kyoto (eds.), *Corpus linguistics: an international handbook*, 899–919. Berlin: Mouton De Gruyter.

Baayen, R. Harald, & Robert Schreuder (eds.). 2003. *Morphological structure in language processing*. Berlin: Mouton de Gruyter.

Beard, Robert. 1998. Derivation. In Andrew Spencer & Arnold M. Zwicky (eds.), *The handbook of morphology*, 44–65. Oxford: Blackwell.

Biber, Douglas. 2009. A corpus-driven approach to formulaic language in English: multi-word patterns in speech and writing. *International Journal of Corpus Linguistics* 14(3). 275–311.

Biber, Douglas, Susan Conrad, & Viviana Cortes. 2004. If you look at . . . : Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics* 25(3). 371–405.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, & Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Longman.

Cheng, Winnie, Chris Greaves, & Martin Warren. 2006. From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics* 11(4). 411–433.

Doğançay, Seran. 1990. Your eye is sparkling: formulaic expressions and routines in Turkish. *Penn working papers in educational linguistics* 6(2). 49–65.

Doğruöz, A. Seza, & Ad Backus. 2009. Innovative constructions in Dutch Turkish: An assessment of ongoing contact-induced change. *Bilingualism: Language and cognition* 12(1). 41–63.

Durrant, Philip. 2009. Investigating the viability of a collocation list for students of English for academic purposes. *Journal of English for Specific Purposes* 28(3). 157–179.

Durrant, Philip, & Julie Mathews-Aydınlı. 2011. A function-first approach to identifying formulaic language in academic writing. *Journal of English for Specific Purposes* 30(1). 58–72.

Ellis, Nick C. 2002. Frequency effects in language processing. *Studies in Second Language Acquisition* 24(02). 143–188.

Ellis, Nick C. 2003. Constructions, chunking, and connectionism: the emergence of second language structure. In Catherine J. Doughty & Michael H. Long (eds.), *The handbook of second language acquisition*, 63–103. Oxford: Blackwell.

Ellis, Nick C. 2008. Usage-based and form-focused SLA: the implicit and explicit learning of constructions. In Andrea Tyler, Yiyoung Kim & Mari Takada (eds.), *Language in the context of use: discourse and cognitive approaches to language*, 93–120. Berlin: Mouton de Gruyter.

Ellis, Nick C., & Diane Larsen-Freeman. 2006. Language Emergence: Implications for Applied Linguistics-Introduction to the Special Issue. 27(4). 558–589.
Fillmore, Charles, J., Paul. Kay, & Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: the case of *let alone*. *Language* 64(3). 500–538.
Gibbs, Raymond W., Nandini P. Nayak, & Cooper Cutting. 1989. How to kick the bucket and not decompose: analyzability and idiom processing. *Journal of memory and language* 28. 576–593.
Göksel, Aslı, & Celia Kerslake. 2005. *Turkish: A comprehensive grammar*. London: Routledge.
Goldberg, Adele E. 2006. *Constructions at work: the nature of generalization in language*. Oxford: Oxford University Press.
Hay, Jennifer B., & R. H. Baayen. 2005. Shifting paradigms: gradient structure in morphology. *Trends in cognitive science* 9(7). 342–348.
Hoey, Michael. 2005. *Lexical priming: A new theory of words and language*. London: Routledge.
Hunston, Susan, & Gill Francis. 2000. *Pattern Grammar: a corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.
Kemmer, Suzanna, & Michael Barlow. 2000. Introduction: a usage-based conception of language. In M. Barlow & S. Kemmer (eds.), *Usage based models of language*, vii–xxviii. Stanford: CSLI Publications.
Laine, Matti, Jussi Niemi, P. Koivuselkä-sallinen, E. Ahlsén, & J. Hynönä. 1994. A neurolinguistic analysis of morphological deficits in a Finnish-Swedish bilingual aphasic. *Clinical Linguistics & Phonetics* 8(3). 177–200.
Langacker, Ronald W. 1987. *Foundations of cognitive grammar: Volume 1 Theoretical prerequisites*. Stanford: Stanford University Press.
Lehtonen, Minna, Toni Cunillera, Antoni Rodriguez-Fornells, Annika Hulten, Jyrki Tuomainen, & Matti Laine. 2007. Recognition of morphologically complex words in Finnish: Evidence from event-related potentials. *Brain Research* 1148. 123–137.
Lehtonen, Minna, & Matti Laine. 2003. How word frequency affects morphological processing in monolinguals and bilinguals. *Bilingualism: Language and cognition* 6(3). 213–225.
Moon, Rosamund. 1998. *Fixed expressions and idioms in English*. Oxford: Oxford University Press.
Niemi, Jussi, Matti Laine, & Juhani Tuomainen. 1994. Cognitive morphology in Finnish: foundations of a new model. *Language and cognitive processes* 9(3). 423–446.
Oflazer, Kemal, Özlem Çetinoğlu, & Bilge Say. 2004. Integrating morphology with multi-word expression processing in Turkish. In Tanaka Takaaki, Aline Villavicencio, Francis Bond & Anna Korhonen (eds.), *Proceedings of the ACL Workshop on multiword expressions: Integrating processing, Barcelona, Spain*, 64–71.
Özkan, Bülent. 2007. *Türkiye Türkçesinde belirteçlerin fiilerle birliktelik kullanımları ve eşdizimliliği*. Adana: Çukurova Üniversitesi PhD thesis.
Pawley, Andrew, & Frances Hodgetts Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In Jack C. Richards & Richard W. Schmidt (eds.), *Language and communication*, 191–226. New York: Longman.
Pawley, Andrew, & Frances Hodgetts Syder. 2000. The one-clause-at-a-time hypothesis. In H. Riggenbach (ed.), *Perspectives on fluency*, 163–199. Ann Arbor: University of Michigan Press.
Peterson, Robert R., Curt Burgess, Gary S. Dell, & Kathleen M. Eberhard. 2001. Dissociation between syntactic and semantic processing during idiom comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 27(5). 1223–1237.

Pilten, Şahru. 2008. Eş Dizimlilik Özellikleri Bakımından Türk Dilinde 'Temiz' Sözcükleri. *Modern Türklük Araştırmaları Dergisi* 5(4). 24–48.
Pinker, Stephen. 1999. *Words and rules: the ingredients of language*. London: Phoenix.
R Development Core Team. 2010. R: A language and environment for statistical computing. In. Vienne, Austria: R Foundation for Statistical Computing.
Sak, Haşim, Tunga Güngör, & Murat Saraçlar. 2008. Turkish language resources: morphological parser, morphological disambiguator and web corpus. In Bengt Nordström & Aarne Ranta (eds.), *Advances in natural language processing: Proceedings of the 6th International Conference on Natural Language Processing, GoTAL 2008, Springer Verlag, LNAI series 5221*, 417–427.
Scott, Mike. 2008. WordSmith Tools 5.0. In. Oxford: Oxford University Press.
Sinclair, John McH. 2004. The search for units of meaning. In *Trust the text: language, corpus and discourse*, 24–48. London: Routledge.
Soveri, Anna, Minna Lehtonen, & Matti Laine. 2007. Word frequency and morphological processing in Finnish revisited. *The mental lexicon* 2(3). 359–385.
Stefanowitsch, Anatol, & Stefan Th. Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2). 209–243.
Stubbs, Michael. 1996. *Text and corpus analysis*. Oxford: Blackwell.
Tannen, Deborah, & Cömert Öztürk. 1989. Health to our mouths: formulaic expressions in Turkish and Greek. In F. Coulmas (ed.), *Conversational Routine*. The Hague: Mouton.
Vartiainen, Johanna, Silvia Aggujaro, Minna Lehtonen, Annika Hulten, Matti Laine, & Riitta Salmelin. 2009. Neural dynamics of reading morphologically complex words. *Neuroimage* 47(4). 2064–2072.
Wray, Alison. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.
Wray, Alison. 2008. *Formulaic language: pushing the boundaries*. Oxford: Oxford University Press.

Bionote

Philip Durrant is Assistant Professor in the Graduate School of Education at Bilkent University in Ankara, Turkey.